Statistiques

1. L'Information statistique

1.1 Qu'est-ce que la Statistique ?

La Statistique est une discipline qui a eu successivement plusieurs objets. La statistique pratiquée par les universités allemandes du XVIIème siècle est proche de la Science politique, et se consacre à la description non chiffrées des choses remarquables des Etats. Elle cherche à comprendre ce qui fait la force de ces Etats. La Statistique anglaise dite Arithmétique politique est née à la même période dans les milieux de la politique et de la finance et vise à établir des relations quantitatives entre des grandeurs comptables et économiques. La statistique Napoléonienne, proche de la géographie, a mené des enquêtes départementales pour établir des monographies descriptives ancrées sur le territoire, ses ressources et les activités humaines qui s'y rapportent.. La Statistique des populations, dite plus tard démographique, cherche à exploiter les enregistrements des naissances et des décès (des flux) et à évaluer la population (stock) de chaque Etat soit à partir de ces flux (méthode du multiplicateur des naissances) soit par recensement. L'institutionnalisation des Bureaux de Statistique un peu partout en Europe au début du XIXème siècle a conduit à une Statistique administrative qui a emprunté à ces différentes traditions, et qui a conquis le monopole étatique de la production de données économiques et sociales de type numérique.

La Statistique a aussi désigné une méthode de réduction des données sous forme de tableaux, de moyennes, et de graphiques (Statistique descriptive, cf. Leçon 2) à partir desquels on a voulu dégager des régularités dans les distributions, dans les liaisons, ou dans les évolutions de quantités observées dans les sciences de l'homme (Arithmétique politique du XVIIème, Anthropométrie, Démographie, Statistique morale, Sociologie, Psychologie, Economie) comme dans celles de la nature (Astronomie, Géodésie, Mécanique, Météorologie). Depuis la théorie des Erreurs développée au début du XIXème par Laplace et Gauss jusqu'à la Statistique mathématique du XXème siècle, le raisonnement statistique s'est attaché dans ces différentes disciplines à la question de l'inférence inductive qui consiste à généraliser à des populations, sous forme de lois, ce qui était observé sur un petit nombre de cas. Cette inférence s'est appuyée sur le calcul des probabilités, né à la fin du XVIIème siècle dans des milieux aussi différent que ceux du commerce, du droit, de la théologie et de l'astronomie, et qui a joué un rôle majeur dans la modélisation du hasard, des sondages, des erreurs et des variations, et dans l'aide à la décision en univers incertain.

Aujourd'hui, la Statistique est considérée comme une discipline scientifique intermédiaire qui n'a plus d'*objet* privilégié (l'Etat, la population, les entreprises, les particules de la physique…) mais qui propose une *méthode*, utilisable par toutes les sciences de la nature et de l'homme, de *production et de traitement de l'information à des fins de connaissance et de décision.* La fin dernière de cette méthode, déjà présente dans l'*Ars Conjectandi* de J. Bernoulli (1713) sous le nom de « stochastique », mais renforcée par les théories de Wald et Neyman dans les années 1930-50, est bien en effet de perfectionner cet art de la conjecture, de la preuve, et de la décision. On peut décomposer cette méthode de la stochastique en quatre phases.

La première phase s'occupe de la construction des faits et de leur mesure et elle a longtemps constitué la seule activité légitime de la discipline. Elle traite de questions aussi fondamentales que celles des catégories, nomenclatures et taxinomies à travers

lesquelles le réel est appréhendé. La Statistique construit des équivalences : *« Les statisticiens ne peuvent compter que ce qui est normalisé par des accords collectifs, conceptualisé par des usages sociaux, questionné par les politiques, conceptualisé par des textes législatifs »* (O. Martin, 1997). On ne peu compter les chômeurs, les cadres supérieurs, les salariés qu'après avoir défini ces catégories. Cette première phase de la Statistique traite aussi des méthodologies de l'investigation, qui comprennent principalement deux catégories : les réutilisations de procédures administratives (par exemple déclaration d'impôt, feuille de sécurité sociale) et les enquêtes statistiques indépendantes de toute gestion administrative. Ces investigations aboutissent à des relevés d'information qu'on appelle parfois des « données » (ou data). Ce terme est, nous le verrons, bien abusif. Cette même phase traite aussi des échelles de mesures sur lesquelles ces catégories sont projetées (cf. infra) pour obtenir une mesure. Cette production d'information est une *construction* sociale et technique des faits par un système d'information statistique (cf. infra Leçon1) qui articule des institutions, des lois et contrats, des définitions des conventions et cadres comptables, des dispositifs d'investigation, de stockage, d'exploitation et de publication.

La seconde phase consiste en un premier traitement des données appelé statistique descriptive (Leçon 2), dont l'objectif est de résumer sans trop la trahir l'information contenue dans une base de données, grâce à des méthodes numériques, tabulaires et graphiques sous lesquelles la même information peut être communiquée. L'information complète sur une mesure d'âge sera remplacée par exemple par une table de fréquence, un histogramme, une moyenne, un écart-type qui donnent une vision synthétique de la distribution.

La troisième phase consiste à produire sur la base des faits contingents un discours théorique général. C'est un moment d'une Statistique que l'on dit parfois inférentielle et qui a pour fonction de fournir les raisonnements, les enchaînements de la preuve, qui prendront place dans un discours argumentaire fondé sur la répétition et les grands nombres. Le raisonnement peut être inductif s'il va des observations aux régularités et aux lois, par exemple si l'on infère d'un tableau statistique que fumer donne le cancer, ou bien de type hypothético-déductif si les observations servent à valider ou réfuter un jeu d'hypothèses a priori qui prend souvent la forme d'un modèle théorique : celui-ci fournit a priori des relations théoriques entre variables que l'on teste sur la base des liaisons statistiques obtenues sur quelques observations de ces variables. La Statistique dans cette phase se situe alors au carrefour des problèmes de logique inférentielle qu'une certaine mathématique peut structurer, et des questions de légitimation sociale que de nombreuses controverses historiques illustrent : on pense aux débats sur l'efficacité de la vaccination, au XVIIIème siècle, sur les causes du paupérisme au XIXème, sur l'explication eugéniste des inégalités au début du XXème, qui ont été les cadres agités dans lesquels certaines formes d'inférences légitimes ont été élaborées.

La quatrième phase est celle de la décision et de l'action, qu'elles soient individuelles ou collectives. Dès l'origine, la Statistique s'est souciée de fournir des règles de décisions optimales dans des questions faisant intervenir des conséquences incertaines : assurances maritimes, rentes viagères, témoignages, décisions de justice sont l'occasion d'élaborer ces règles et de les fonder sur un calcul nouveau, celui de la valeur (ou l'utilité) espérée. Le XXème siècle a vu ressurgir cette problématique de l'action dans le domaine de la recherche opérationnelle, de la théorie des choix collectifs, du choix des investissements, de la détermination d'une politique économique optimale. La Statistique est alors un lieu de tension entre des logiques scientifiques et des logiques politiques, puisqu'elle aide à définir ce qui est légitime [Brian, 1994 ; Desrosières, 1993].

Ces phases articulent quatre formes de la *raison statistique* : la première se rapporte à la définition et à l'*existence* même des *faits* ; la seconde traite de leur mise en *forme* numérique par des opérations syntaxiques de mise en série ou d'agrégation ; la troisième est de l'ordre du *discours* et de la sémantique ; elle construit des représentations significatives de leur organisation, et traite de leur *pertinence*. La quatrième est de l'ordre des *actes* et de la pragmatique ; elle fournit des règles d'action individuelle ou collective, et traite de leur *efficacité* par rapport aux objectifs visés et de leur *valeur* par rapport à tel ou tel système éthique.

L'économiste, le sociologue, le gestionnaire sont le plus souvent engagés dans des études qui nécessitent la connaissance de la méthodologie statistique. Pour établir *des faits stylisés*, ou pour estimer les valeurs des paramètres du *modèle* qu'ils construisent pour tester une théorie, ou plus souvent pour évaluer des variantes de politique économique, pour faire des *prévisions*, pour prendre une *décision*, ils doivent avoir une bonne connaissance du système statistique qui produit les « données » à mobiliser ; ils doivent connaître les règles mathématiques de l'inférence statistique qui permettront de valider leurs modèles théoriques ; il doivent savoir utiliser ces résultats dans des expertises et des aides à la décision en matière de politique économique et sociale pour les uns, en matière de gestion de l'entreprise pour les autres.

1.2 Données

La principale activité de nos sociétés, y compris en liaison avec la production de biens et services, est la gestion de l'information attachée à cette activité primaire. Cette information prend des formes très variées : son contenu peut concerner des *états* (état économique, politique, social d'une nation, bilan d'une société, état d'un patrimoine.....), des *faits* (événements, flux d'activités, mouvements de biens et de personnes), des *discours.* Sa forme dépend de la trace ou de la représentation qui est conservée de ces contenus. L'information peut prendre la forme matérielle de *textes*, de *nombres*, d'*images fixes ou animées,* de *sons*, ou d'un mélange de ces formes. L'information circule sur des *supports* matériels (objet, livre, journal, revue, film.) ou immatériels (ondes, impulsions, bits) qui sont autant de *média,* appartenant à des systèmes médiatiques comme l'imprimerie, la presse, la radio, la télévision, l'affichage urbain, la télématique.

Quel que soit son contenu, sa forme, ou son support, l'information pour être stockée ou échangée, doit être découpée en unités plus ou moins grosses : l'unité de base est une « donnée », par exemple le montant des ventes de tel produit par telle entreprise pendant telle année sur tel marché. Les « *données* » sont regroupées par *fiches* ou *enregistrements* concernant une même unité statistique, dans des *fichiers*, eux-mêmes regroupés en *banques de données*, ou datawarehouse, et à une échelle plus grande, en réseaux de banques de données.

Les « données » portent très mal leur nom. Elles ne sont pas « données », aux deux sens (épistémologique et financier) de cette expression, puisqu'elles sont *produites* par un *système d'information statistique* et que, de ce fait, elles ont un *coût*. D'un point de vue épistémologique, les objets de science ne sont jamais des « données », mais des reconstructions conceptuelles ou opérationnelles faites par les chercheurs et les institutions. Le « réel » de l'économiste ou du gestionnaire ne se donne pas à lire, il est construit par son jeu d'hypothèses, de concepts, et d'outils, et par ses objectifs ; mais il est également instrumenté par le système d'information auquel il emprunte ses « données ». La consommation totale, le revenu national, le nombre de chômeurs ne sont pas des « données », mais des grandeurs dont la définition, la mesure, la compatibilité avec d'autres grandeurs (Comptabilité nationale), les opérations qui conduisent à des estimations, et les négociations et arbitrages entre sont garanties par certaines institutions : on parle ainsi du chômage au sens du BIT, des cadres au sens de la classification des PCS, des entreprises de transport au sens de la nomenclature européenne des activités…etc. Les entreprises ont également besoin de définir des nomenclatures de leurs produits, de leurs employés, et de leurs activités, et pour des raisons contractuelles, ou réglementaires et légales, elles doivent le faire dans ces cadres institutionnels. La définition et la mesure de la plupart des grandeurs sont fixées et garanties socialement. D'autre part la production de ces « données » est une opération en général fort coûteuse même si ces données deviennent des biens et des services publics. De plus en plus cependant elles entrent dans des circuits marchands, et ont un prix. Le coût global des opérations statistiques réalisées par l'INSEE et les services statistiques des ministères était de 1 672 millions de francs en 1985 [Ousset, 1992].

Avant de travailler sur des « données », il convient donc de s'interroger sur les conditions de leur production. Qui les a produites, avec quel statut, dans quel objectif, pour répondre à quelle demande, avec quel dispositif, sous quelle forme, avec quelle disponibilité ?… En bref il faut connaître le système d'information qui est leur matrice.

Toutes les « données » de l'économiste ou du gestionnaire sont le produit d'une demande et d'une offre d'information, qui peuvent d'ailleurs ne pas être en phase. Il y a parfois conflit d'intérêts entre le producteur du système d'information et l'utilisateur de ses produits qui n'y trouve pas de quoi satisfaire ses besoins, surtout dans le cas du monopole. L'analyse critique des sources est d'ailleurs une partie de l'art de l'utilisateur de statistiques. Celui-ci doit donc savoir *localiser, consulter, et critiquer* les sources d'information statistiques propres à son domaine.

Aux publications de l'INSEE, des Administrations (Ministères), et des Instituts de sondage ou de recherche (recensements, comptes de la nation, résultats d'enquêtes) sur support papier, il faut ajouter les ressources des nouveaux supports que sont les CD-ROM et les banques de données en ligne sur Internet. Voir à ce sujet le catalogue des ressources électroniques de la BU de l'Université et les banques de données mises à disposition des étudiants distants par CANEGE.

1.3 Système d'information

Le système d'information statistique est l'ensemble des éléments hétéroclites (des pratiques, des méthodes, des institutions) interdépendants, matériels et immatériels, formels et informels, qui concourent à la production d'information dans un champ donné [Michel Lévy, 1975]. Autre définition de Michel Volle : Ensemble constitué par la définition des processus des métiers et par celle des stocks et flux d'information éclairant ces processus. Un système d'information peut être *privé* c'est à dire propre à une organisation, et lui permettant de prendre toute décision. Le système d'information d'une entreprise par exemple rassemble des informations sur ses opérations internes (production, administration...) aussi bien que sur son environnement externe (le marché, la réglementation, la concurrence…) et permettent un management éclairé. Le système d'information public agit de même pour le compte des administrations publiques. Comment par exemple connaît-on quelque chose de l'activité des entreprises françaises ?

Pour y arriver, il faut cumuler et combiner une dizaine d'éléments qui forment le système d'information statistique sur la production industrielle française. Un tel système s'est réellement mis en place dans les années 1960, même si des éléments de ce système datent des deux siècles précédents. Faisons la liste de ces éléments du système :

1. Il faut d'abord *une volonté politique* de produire cette information et de la mettre au service des entrepreneurs et d'un public plus ou moins large [Michel Volle, 1982]. Toute l'information sera conditionnée par cette volonté politique et par les motifs - nobles ou moins nobles - qui l'inspirent : connaissance des mécanismes, action sur ces mécanismes, optimisation de ceux-ci, mais aussi publicité, répression, manipulation. De plus cette volonté politique doit s'accompagner d'un désir démocratique de partager cette information en la rendant publique. Le système public d'information statistique est profondément lié à la démocratie, et si l'ancien régime produisait des statistiques, celles-ci restaient le plus souvent le fait du prince (du roi) et de ses intendants et réservées à leur usage, avec rare publication (souvent à l'étranger) et diffusion sous le manteau. Même sous un régime démocratique, les relevés statistiques n'existent que parce que le problème qu'ils éclairent est « mis à l'agenda » des politiques, reconnu d'utilité publique, et doté de nouveaux moyens d'investigation.

2. Il faut des *institutions* qui rendent opérationnelle cette volonté politique, et qui sont chargées de produire et diffuser cette information. L'INSEE, premier producteur public, créé par la loi du 27 avril 1946 en relais de la SGF et du SNS, a pour vocation la collecte des statistiques, l'étude de la situation économique, la coordination des méthodes et des travaux des administrations publiques, la diffusion ou publication des résultats, la formation du personnel, et la recherche-développement en statistique et économie. Les services statistiques ministériels, les organismes de recherche publics (CREDOC, INED, CEREQ…) ou privé (IFOP…), et, au niveau européen ou international, les officines comme Eurostat, l'OCDE, le BIT complètent le dispositif.

3. Il faut des accords et *contrats* entre ceux qui détiennent l'information et l'autorité publique qui souhaite la recueillir, l'agréger et la diffuser, voire des *lois* comme celle qui réglemente l'obligation et le secret statistique : la loi du 7 juin 1951 crée l'obligation légale de répondre aux questionnaires contre la garantie du secret c'est à dire d'une utilisation purement statistique de l'information, et introduit un organe de consultation et de coordination, le CNIS. La loi sur l'informatique et les libertés (6 janvier 1978) a complété le dispositif légal (demande d'autorisation pour tout fichier, et limitation des croisements de fichiers).

4. Des *nomenclatures* définissent, par convention et arbitrage plutôt que logiquement, les catégories ou classes d'objets ou d'individus sur lesquels s'exerce la mesure. On ne connaît rien de chacun des millions d'objets produits, mais seulement de certains groupes d'objets. La nomenclature d'activité et de produit (NAP) en vigueur en France de 1973 à 1992 a été remplacée par la NAF qui est une adaptation de la NACE européenne. L'ONU de son côté a adopté une classification internationale type (CITI). La nomenclature des professions et catégories professionnelles (CSP) a été refondue en 1982 (PCS). Des cadres comptables (le SECN) définissent les grandeurs à mesurer et leurs liens.

5. Des *instruments d'investigations* sont ensuite nécessaires pour recueillir l'information. Ils forment le noyau méthodologique d'un système d'information. A l'exploitation toujours possible mais biaisée de procédures administratives comme l'imposition sur les Bénéfices Industriels et Commerciaux (BIC), on préfère soit des recensements, soit (moins coûteux) des enquêtes périodiques comme l'enquête annuelle d'entreprise, ou l'enquête de conjoncture, enquêtes spéciales à une branche ou un champ.

6. Des *techniques de codage, stockage et traitement de l'information* complètent les moyens d'investigation pour inscrire l'information dans des supports non volatiles : hardware (machines mécanographiques puis ordinateurs) et software (systèmes de fichiers et bases de données). Le répertoire national d'identification des personnes physiques (RNIPP) qui remonte à 1940 code les individus par un numéro national d'identité à 13 chiffres. Le répertoire des personnes physiques non salariées ou des personnes morales qui relèvent du registre du commerce et des sociétés ou du répertoire des métiers, est constitué sous la forme d'un fichier SIRENE des entreprises et des établissements.

7. Des moyens de *diffusion et publication* sont offerts en retour aux agents ayant contribués à cette information aussi bien qu'aux autres citoyens : supports imprimés (annuaires, revues, catalogues) ou électroniques (off-line ou on-line). C'est l'aspect visible de l'iceberg du système d'information, que l'on appelle les *sources*, et leur conditions d'accès (publiques / privées, gratuites / marchandes).

8. Des moyens de *recherche* et de *formation* autour de la technologie de l'information (Universités, Ecoles, Laboratoires) visent à perpétuer et améliorer les éléments de ce système et son organisation générale.

Nous allons reprendre plus en détail la question des méthodes d'investigation (point 5) qui constituent la pièce centrale d'un dispositif d'information.

1.4 Enquêtes

L'information statistique est le produit d'un système statistique public (étatique) ou privé (de l'entreprise ou l'organisation). Agenda politique, lois, conventions, nomenclatures définissent le cadre de cette production. Mais comment se fait la production elle-même ? L'information peut être produite de deux façons :

Comme produit dérivé d'une activité administrative. C'est le cas des revenus donnés par l'impôt, des dépenses de santé connues par les caisses d'assurance-maladie…L'avantage est l'utilisation d'une information déjà recueillie à des fins d'administration. Mais le biais des résultats est important parce que le champ souhaité de l'enquête, c'est à dire la définition des populations concernées et des mesures effectuées, ne coïncide pas en général avec celui de l'activité administrative. Les ménages imposés ne sont qu'une partie des ménages, et les revenus déclarés au fisc ne sont pas les revenus économiques. Les consommations médicales (consultations, médicaments, hospitalisations) connues des organismes de gestion de la Sécurité sociale ne sont qu'une partie des consommations, celles qui donnent lieu à prise en charge. Il est donc indispensable de concevoir des dispositifs dont l'objectif est uniquement la production d'information statistique.

L'*enquête* est la forme commune de ces dispositifs. Le mot qui désigne aussi bien l'enquête policière que l'enquête administrative ou l'enquête scientifique a la même origine que le mot inquisition : il s'agit bien de faire parler les détenteurs d'information. Un des détournements courants de l'enquête administrative ou privée est de ne pas identifier les enquêtés comme porteurs d'information mais comme cible d'une information dirigée qui ne dit pas son nom : c'est le cas de trop nombreuses pseudo-enquêtes de « communication » qui servent de camouflage à des opérations de persuasion et manipulation. Même si l'on exclut ces cas de détournement à d'autres fins d'une opération de connaissance, l'enquête est déjà un coup de force qui prend acte d'une dissymétrie de pouvoir entre l'enquêteur et l'enquêté, le premier étant en mesure d'imposer un protocole au second et de lui soutirer une information qui l'intéresse, en général sans contrepartie, sauf dans le cas d'un contrat explicite.

L'enquête peut être ponctuelle, occasionnelle, ou régulière (enquête annuelle d'entreprise). On distinguera d'abord les *enquêtes intensives*, ou monographiques, dont le modèle est la technique de Le Play, un « polytechnicien sociologue » du milieu XIXème, et qui peuvent prendre la forme de l'observation participative par immersion prolongée des ethnologues, de l'entretien plus ou moins directif des journalistes et psychologues. A ces formes intensives riches mais non cumulables, on oppose les *enquêtes statistiques extensives* sur de larges populations, qui permettent d'accéder à la généralité des énoncés par inférence inductive contrôlée, ce que ne permet pas facilement la monographie.

Les enquêtes peuvent être exhaustives (elles s'adressent à toute la population) comme c'est le cas du recensement périodique de la population française. Mais elles ont alors un coût prohibitif. Ou bien ce sont des enquêtes par *sondage* qui ne s'adressent qu'à une partie de la population appelée *échantillon*. La question qui se pose alors le plus souvent est d'assurer la représentativité de l'échantillon par rapport à la population pour permettre ensuite le raisonnement inférentiel. Ce n'est pas avant 1925 que le sondage a été reconnu par les statisticiens de l'IIS comme méthode d'investigation admissible. Et son succès public date du sondage de Gallup à l'élection présidentielle américaine en 1936.

La représentativité de l'échantillon peut être assurée de deux manières. Par *choix raisonné*, on s'impose seulement la proportionnalité des effectifs de l'échantillon et de la population sur les principales variables indépendantes (sexe, habitat, CSP…) : c'est par exemple la méthode des quotas. Les individus peuvent ensuite être choisis de façon quelconque, par exemple la plus économique. La méthode du *sondage aléatoire* suppose

par contre 1°) que l'on dispose de la liste des unités de la population (la base de sondage) 2°) Que l'on tire au hasard (avec une table ou un ordinateur) les individus de l'échantillon. Tirage au hasard veut dire avec des probabilités connues, qu'elles soient égales ou inégales, et cela peut se faire à plusieurs niveaux, avec ou sans stratification. Seule la méthode aléatoire permet ultérieurement d'utiliser le calcul des probabilités pour chiffrer l'erreur d'échantillonnage (Voir Leçon 5).

L'enquête par questionnaire est la forme la plus courante de l'enquête extensive. La construction d'une enquête passe par les phases suivantes (cf. schéma du cycle de l'enquête page suivante) :

1. Le champ de l'enquête est le domaine (sujet et population concernée) sur lequel elle porte : par exemple le chômage et les chômeurs. L'objet est une reconstruction conceptuelle de ce domaine : le chômage peut être abordé par des concepts psychologiques, sociologiques, économiques…Une problématique est un ensemble de questions logiquement reliées entre elles que l'on se pose sur l'objet (les origines des chômeurs, leur indemnisation, leur insertion, leurs loisirs…). Il est nécessaire d'avoir au départ quelques hypothèses à tester sous peine de questionner les enquêtés sur tout et n'importe quoi. Le champ, l'objet, la problématique, et les hypothèses d'une enquête sont négociés entre le chercheur, l'organisme et le commanditaire.

2. La construction d'une enquête suppose d'abord une analyse conceptuelle qui décompose la l'objet en plusieurs dimensions. Par exemple une enquête sur l'autonomie des étudiants s'intéressera successivement à son autonomie financière, matérielle, affective, idéologique…Mais on ne peut directement demander à un étudiant s'il est autonome ! Chacune de ces dimensions devra donc ensuite être traduite en quelques indicateurs observables qui remplaceront une abstraction (indépendance matérielle) par des items relatifs au statut, au comportement, aux opinions (où loge l'étudiant, où mange-t-il, qui lave son linge, vote-t-il comme ses parents…).

3. Chacun de ces indicateurs est ensuite traduit en une question, fermée (nombre fini de modalités) ou ouverte. La formulation de questions est un art subtil dont les effets sur les réponses sont importants. L'assemblage de ces questions en un questionnaire structuré est tout aussi délicat car la proximité de deux questions, leur ordre, leur articulation, modifient leurs interprétations par l'enquêté, et donc influencent les réponses.

4. Le questionnaire peut être administré sous différentes formes : en face à face, par dépôt de carnet, par correspondance, par téléphone...L'interaction entre enquêteur et enquêté doit se faire dans des conditions d'identité parfaite et de neutralité pour que les réponses de deux enquêtés soient comparables. D'où la nécessité d'un protocole fixe, explicité, le même pour tous, que devront respecter les enquêteurs, même si les situations sont variables. C'est la condition de « mise en équivalence » qui permettra ensuite de dire qu'une réponse en vaut une autre, est homologue à une autre et donc sommable et dénombrable. Le remplissage d'un questionnaire n'est pas un simple enregistrement, c'est une interaction stimulus-réponse. Une opinion n'est pas recueillie, elle est sollicitée voire provoquée. On retrouve une certaine violence symbolique dans l'imposition d'une démarche, d'un raisonnement, de présupposés que vous avez certainement ressentis en tant qu'enquêté.

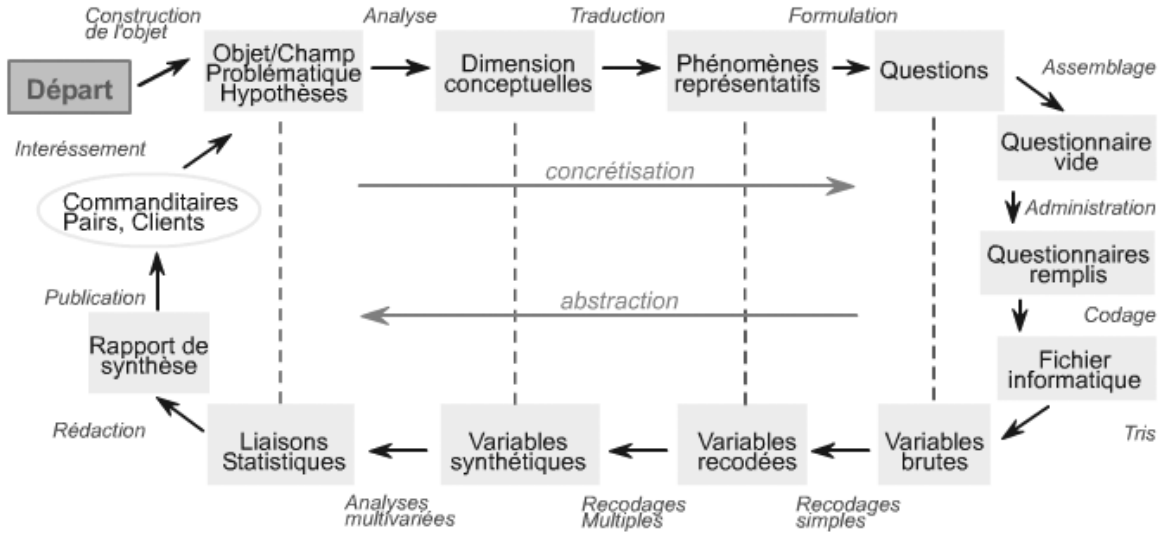
5. Les questionnaires remplis sont ensuite soumis à un codage des réponses et une saisie qui conduisent à la production d'un fichier. Ce codage, nécessaire à un traitement qui a été historiquement manuel, puis mécanographique, puis informatique, est à la fois une convention à établir (qui n'est pas toujours neutre pour les inférences qui seront faites plus tard) pour qu'une modalité de réponse deviennent un code, et une opération pratique qui transforme chaque modalité en un code inscrit dans un *champ*, et chaque questionnaire en un *enregistrement* constitué de la suite de ces champs.

6. Le dépouillement d'une enquête consiste en une analyse statistique de ce fichier sous la forme principale de tris simples et multiples. Un tri simple (ou tri à plat) ne produit qu'une information brute et peu intéressante caractérisant la distribution des individus sur les modalités d'une variable. A l'occasion de ces tris, on doit très souvent procéder à des recodages des variables : regroupement de modalités, construction d'indices synthétiques.

7. Les tris doubles ou multiples produisent des relations entre les variables (associations sous forme de tableaux croisés, liaisons sous forme de relation fonctionnelle, analyses factorielles…) qui sont d'un plus grand intérêt.

8. L'enquête se clôt par un rapport de synthèse à destination des commanditaires qui produit des interprétations de ces résultats statistiques, sous forme d'énoncés dans la langue commune et scientifique, et qui vise à répondre aux questions de la problématique de départ.

Le schéma ci-dessous traduit la succession des phases d'une enquête par un cycle à quatre temps : concrétisation de la problématique en un protocole, administration de ce protocole, abstraction par traitement statistique des données, synthèse et interprétation des relations mises en évidence par le traitement.



Cycle de l'enquête par questionnaire

1.5 Modèles

La méthodologie de l'enquête ainsi décrite, dans laquelle on part des observation pour en inférer des régularités, des relations ou des lois, est principalement de type *inductive*, même si l'ensemble de départ objet-problématique-hypothèse suppose une approche théorique préalable qui restreint les risques d'une induction libre. Dans une approche plus « falsificationniste » (selon Popper), les économistes préfèrent formuler (spécifier) a priori un modèle théorique qui exprime par des équations les relations comptables et fonctionnelles entre des variables exogènes (explicatives) et des variables endogènes (expliquées). Depuis les travaux de la Cowles Commission dans les années 1940, ce sont des modèles structurels stochastiques (avec un terme aléatoire dans chaque équation de comportement) qui traduisent les comportements des agents économiques. Les économistes limitent alors leur investigation statistique aux seules variables qui permettent d'estimer les paramètres de ce modèle et de tester certaines hypothèses qui lui sont liées.

Les quatre phases principales de la modélisation économique sont alors les suivantes :

1. Spécification du modèle, c'est à dire écriture sous forme d'équations de définitions et de comportement des principales hypothèses concernant le système économique décrit, et choix des propriétés de la distribution des termes aléatoires du modèle.

2. Estimation statistique des valeurs des paramètres inconnus du modèle à partir des séries de valeurs observées sur les variables exogènes et endogènes.

3. Etude des propriétés dynamiques du modèle.

4. Utilisation en prévision ou en simulation de variantes.

Depuis la fin des années 1980, cette modélisation structurelle propre à la méthode économétrique a connu d'importants bouleversements : remise en cause de la distinction faite a priori entre variable exogène et variable endogène, fondement de la macroéconomie sur un comportement microéconomique des agents compatible avec un principe d'équilibre général et susceptible d'anticipations sous la forme adaptative puis rationnelle, meilleure prise en compte des processus temporels dans le traitement des séries observées (stationnarité, autocorrélation, cointégration).

A ces modèles conceptuels qui font encore une part importante à la statistique, soit par les bases de données qui permettent un calage du modèle, soit par une référence à des procédures d'optimisation stochastiques explicites, on peut opposer des modèles globaux et empiriques de simulation de systèmes complexes (transport, écosystèmes, climat…) dans lesquels l'objectif n'est plus de traduire ou valider une théorie (dont on n'a pas tous les éléments) mais de simuler le comportement dynamique futur d'un système décrit par des relations quantitatives entre des inputs et des outputs, entre des flux et des stocks de toute nature. Les données statistiques fournissent alors une base de calage et de couplage de ces sous-systèmes, mais dans des simulations de long terme il n'est plus possible de leur faire jouer un rôle de validation théorique.

1.6 Typologie et forme des données statistiques

Les données statistiques peuvent se présenter sous des formes très différentes qu'il faut savoir reconnaître.

Données individuelles : La forme la plus structurée qui a la faveur du statisticien, et qui résulte directement d'un codage d'enquêtes par questionnaires, est la base de *données individuelles* dans laquelle les données prennent la forme d'un *tableau* (auquel correspond matériellement un fichier informatique) dans lequel chaque *individu* (ou *unité statistique*) est une ligne (une fiche ou un *enregistrement* dans le fichier), et chaque colonne une *variable.* A l'intersection de la ligne *i* et de la colonne *j* se trouve la *valeur* ou *modalité* x_{ij} de la variable *j* pour l'individu *i*. Du point de vue informatique, cette modalité occupe le jème *champ* du ième *enregistrement*. Vous en trouverez un exemple avec le fichier *Dau.xls* dont chacune des 621 lignes représente le vecteur (ou protocole) des réponses d'un étudiant à 76 questions.

Variables→ Individus↓	X_1	X_2	…	X_j	…	X_p
1						
2						
……						
i				x_{ij}		
……						
n						x_{np}

Données agrégées : La base des données individuelles qui résulte d'une enquête est rarement disponible, si bien que les données publiées ont le plus souvent la forme de *données agrégée*. L'agrégation peut se faire au niveau des individus de I : on perd alors l'information individuelle et l'on ne conserve que les fréquences simples (absolues n_{jk} ou relatives f_{jk}) des individus qui ont la valeur *k* pour la variable *j* (tri simple), ou les fréquences conjointes sur plusieurs variables (tris multiples). L'agrégation peut se faire au niveau des variables : on n'accèdera qu'à des cumuls ou des combinaisons de variables comme la consommation tous biens confondus. Elle peut encore se faire au niveau des modalités de la variable regroupées par classes de valeurs ou de modalités, ou ramenées à leur valeur centrale. En combinant la première et la dernière forme d'agrégation, on donnera par exemple la répartition du revenu par classe (ou même le revenu moyen) des ménages par type de ménage, par branche, par région…

Coupes, séries, panels : La répétition des observations d'une même grandeur peut se faire dans l'espace ou dans le temps : si l'on observe n unités statistiques distinctes (individus, ménages, entreprises, nations...) au même moment, la suite $\{x_{ij}, i = 1,n\}$des valeurs observées d'une variable j forme une *coupe instantanée,* ou encore *une série transversale*, ou encore une « *cross-series ».* Si au contraire on observe et mesure une même unité statistique (le plus souvent un agrégat au niveau d'une région, d'un secteur, ou d'une économie nationale) à des dates successives, en général périodiques (obs. quotidiennes, hebdomadaires, mensuelles, trimestrielles, annuelles), la suite $\{x_t, t = 1,T\}$ des valeurs observées forme une *série chronologique*, une *série longitudinale*, ou « *time-series* ».

1.7 Type algébrique des variables

Le type des données ou plus exactement le *type algébrique des variables* résulte du type de mesure que l'on a effectué sur les unités statistiques. Une mesure est, du point de vue mathématique, une application de l'ensemble de ces unités dans une structure algébrique, plus communément appelée une échelle.

a) Si cette structure est celle d'une *partition* de l'ensemble des valeurs en classes d'équivalences, telle qu'un individu est affecté dans une classe et une seule, alors la variable est dite qualitative, ou plus précisément *nominale.* L'ensemble (dit ensemble-quotient) des classes est appelé une nomenclature*.* Il n'y a pas d'autre relation entre ces classes que l'exclusion mutuelle. Les modalités « célibataire », « marié », « divorcé », « veuf » de la variable « Etat civil » forment une telle variable nominale. Les modalités {rouge, bleu, vert, jaune, noir…} d'une couleur constituent celle-ci en une variable nominale. Si l'on « mesure » la catégorie socioprofessionnelle d'un individu selon les catégories décimales (codées 0,1,2,…,8) de la nomenclature PCS 1982 (Cf. Desrosières et Thévenot 1988) on n'obtient pas une variable numérique (les chiffres ne sont que des codes et pas des nombres ayant la propriété que 6 soit égal à 3 fois 2), mais une variable nominale.

Un cas particulier de variables nominales est celles qui n'ont que deux modalités, ces *variables indicatrices* d'une propiété P valent 1 si la propriété est vraie, et 0 sinon. On les appelle aussi variables *de Bernoulli,* ou *variables logiques*, ou variables de *présence-absence*, ou variables « *dummy ».*

Une variable nominale a p modalités peut être recodée en p variables indicatrices. Par exemple si l'état d'un bien est caractérisé par les modalités « mauvais », « passable », « bon », la variable qualitative ainsi constituée peut être recodée disjonctivement en trois variables indicatrices M, P, B prenant l'une la valeur 1 et les deux autres la valeur 0.

b) Si on peut définir sur cet ensemble de classes une relation d'ordre total, c'est à dire être capable d'ordonner de façon unique et consensuelle toutes les modalités de la plus « faible » à la plus « forte », alors la variable est dite *ordinale*. C'est le cas par exemple des modalités de réponse « pas du tout », « pas tout à fait », « un peu », « beaucoup », « énormément » à une question du genre « aimez vous A » ou « êtes vous d'accord avec A ». C'est encore le cas de la suite des diplômes Bac, DEUG, Maîtrise, DESS. Mais cela ne l'est plus pour la suite de modalités Bac, DUT, BTS, DEUG, Maîtrise, DEA, DESS dans laquelle il n'y a pas consensus sur l'ordre des 3 diplômes de Bac+2 ou des deux diplômes de Bac+5.

c) Si cet ensemble est muni d'une origine et d'une unité et que l'addition y prend un sens alors la mesure devient quantitative ou numérique et la variable peut être dite *cardinale*. Si la structure numérique qui porte les résultats de la mesure est discontinue, à savoir qu'entre deux modalités quelconques il n'y a qu'un nombre fini de modalités possibles, alors la variable est dite *discrète.* C'est le cas très commun des nombres entiers naturels (N) ou relatifs (Z), par lesquels on mesure un nombre d'enfants, un stock de marchandises, un nombre de transactions etc…

d) Si au contraire il existe un nombre infini (dénombrable ou non) de modalités possibles entre deux modalités numériques, ou dit que la variable est cardinale et *continue,* comme c'est le cas si la mesure est une projections dans l'ensemble des nombres rationnels (Q, qui s'expriment par un rapport de deux entiers) ou dans

l'ensemble des nombre réels (R). Dans chacun des deux cas, la multiplication et la division sont deux nouvelles opérations possibles, mais c'est seulement dans R que se trouvent aussi les irrationnels et la « puissance du continu ».

Cette typologie grossière des variables et des échelles de mesure qui y sont associées est importante car selon le type de la variable, certaines opérations statistiques (le cumul des fréquences, la somme des modalités…) sont ou ne sont pas définies, certains résumés ont ou n'ont pas de sens, et les modèles probabilistes sous jacents seront également différents, comme nous le verrons dans les prochaines leçons. Mais notons bien que ceci caractérise la mesure et pas la grandeur : une même grandeur comme l'âge peut être mesurée par une variable nominale (les jeunes et les vieux), ordinale (les positions dans le cycle de vie), numérique discrète (en années révolues) ou numérique (quasi)continue (en années, mois, jours, heures et minutes). La précision (et la richesse des propriétés) y est croissante mais de plus en plus illusoire, car la signification en est décroissante.